基因組資料庫的開發

＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊＊ ◎農試所農藝系　王子明　王強生

一、前　言

人類基因組是目前國際性的熱門研究課題，目的在於繪製更詳盡人類基因組的遺傳及物理圖譜，以及展示更詳盡的人類基因組DNA 序列。遺傳圖譜可以標示基因或遺傳標誌在染色體上的線性排列，主要的遺傳圖譜有：連鎖圖譜(linkage map)及物理圖譜(physical map)兩種。連鎖圖譜中遺傳標誌間的距離代表著減數分裂時標誌間的重組頻度。物理圖譜中遺傳標誌間的距離則代表位於標誌間的核苷酸數目。

人類基因組的研究具有六個主要目標：(一)建構高解析度的遺傳圖譜；(二)找出物理圖譜的變異處；(三)決定DNA的完整序列；(四)分析類似人類模式生物的特定染色體組；(五)發明自動繪圖技術，以繪製完整基因組的遺傳圖譜、物理圖譜及DNA序列；(六)發展計算工具，如程式、軟體及資料庫，以利收集、說明及傳播人類基因組龐大複雜的圖譜及序列資料。

遺傳圖譜可作為搜尋遺傳疾病基因的可靠來源，並進一步研究此等基因的構造、功能及表現。因此，新一代的高解析度的人類基因組遺傳圖譜能使醫學領域長遠進步；如遺傳諮詢、預測遺傳疾病的罹患率、診斷測試、基因治療、理論的藥物設計(rationale drug design)及藥理遺傳學上的藥物訂製(pharmacogenetic drug customization)等。基因組的研究會產生大量資料，於是美國國家衛生研究院(NIH)搜集公開發表的DNA序列，集合成為分子資料庫，即為genbank。在公元二千年二月公布的genbank分子資料庫包含 5,691,000個 DNA 片段，即等於近5,805,000,000 DNA。人類基因組研究的當前目標是發展新一代的 DNA 定序技術，進行基因組 DNA 定序，定序的速度以指數級數的速度成長，甚至超過 DNA 的合成速度。

運算分子生物學(computational molecular biology)，即以數學計算的方式分析生物的大分子。依據分子生物學中心教條(centra dogma)，可將遺傳資料劃分為：基因組學(genomics)，基因表現(gene expression)及蛋白組學(proteomics)等三個分析層次，以闡明人類基因組圖譜中的構造性及功能性，進而分析人類基因組資料庫。資料庫的開發(database mining)就是從有限的資料中發掘，獲得有用的資訊。雖然繪製人類基因組

作者：王副研究員強生
聯絡電話：(04)23302301轉137

圖譜的計算問題，可以藉由超級電腦來解決，但圖譜的繪製仍然爲醫學研究上的一大挑戰。本文主要以人類基因組爲範例，說明基因組資料庫的開發情形，然而，此一模式將可以應用至所有生物的基因組研究。

二、基因組學(Genomics)

基因組學係以系統性分析基因組，如生物體完整染色體組數及基因組數，可分爲構造組學及功能組學；構造組學係依據 DNA 序列及基因圖譜所組成，並由已知結構 DNA 序列的功能分析，獲得功能基因組學的組成資料。雖然目前人類基因組研究著重於構造組學的研究，但在基因組 DNA 定序工作接近完成時，人類基因組的研究即將進入動態的功能性研究。

基因組的資料可分爲四部份，即 DNA 序列、物理圖譜、遺傳圖譜、參考目錄，應用於貯存和分析遺傳及物理圖譜。DNA 序列資料應包含解析過的分子序列。物理圖譜資料包含八個領域：序列標誌位置(sequence-tagged site)、編碼區域(coding regions)、非編碼區域、控制區域(control regions)、端粒(telomere)、著絲點(centromere)、重複片段(repeats)及中期染色體條帶(metaphase chromosome bands)。遺傳資料包含七個領域：基因座(locus)名稱、基因位置、重組距離(recombination distance)、多型性(polymorphisms)、斷裂點(breakpoints)、重組(rearrangements)及與疾病關聯性(disease association)等。參考目錄應著重於科學及醫學文獻的引用。

基因組資料依據內容可分爲四類，即分子、遺傳、生物體、目錄。分子資料具有四組代表性工具：

1.European Molecular Bio-logy Laboratory Nucleotide Sequence Data Library (EMBL): http://www.embl-heidelberg.de/
2.DNA Database of Japan (DDBJ)： http://www. ddbj.nig.ac.jp/
3.Genbank16: http://www.ncbi.nlm.nih.gov/Genbank /GenbankSearch.html
4.Swiss Prot: http://www.expasy.ch/sprot/sprot-top.html。遺傳資料工具有：Genome Database (GDB)：http://www.gdb.org與Online Mendelian Inheritance in Man (OMIM)：http://www. ncbi.nlm.nih.gov/Omim/。

生物體資料工具有：Bacterium *Escherichia coli*. Mouse *Mus musculus*.及Mustard plant *Arabidopsis thaliana*.。目錄資料工具有：Biological Abstracts、CancerLit、Excerpta Medica (Embase)及Medline。

三、資料庫的開發 (Database Mining)

基因組資料庫是目前積極開發的資料庫，而資料庫開發的方法係由解析基因組的運算結果而來，目前正使用的方法是利用由有關基因組功能基因的已知位置及 DNA 序列來驗證未知的 DNA 序列的功能。解析基因組的運算可分構造性的解析及功能性的解析二個連續步驟。構造性解析利用基因搜尋程式(computational gene discovery algorithms)來尋找 DNA 序列中基因的開放編碼區(open reading frames, ORFs)。功能性解析則是利用已知基因功能的序列，進行相似性比對，以預測目標基因的功能。當 DNA 序列中重要的功能性位置被陸續研究出來，部份特性即以型式辨識程式(pattern-recognition algorithms)進行分析。因此，一旦 DNA 功能位置的序列完成定序，並找出專一結合的

蛋白質，例如啓動子(promoter)，型式辨識程式隨即具有發現正確位置與排除錯誤位置等兩種能力。

當 RNA 自基因組 DNA 轉錄時，隱子一顯子(intron-exon)間之界線及接合位置的確認是非常重要的。所以具有預測隱子位置的能力時，可以大幅度降低將基因組的 DNA 序列直接轉譯爲胺基酸序列的錯誤發生。而未知基因的 DNA 序列通常與已知基因有構造上的相同性，因此，在比較生物間的親緣關係上，DNA 序列的比較是一種重要的技術。多序列比對分析(multiple sequence alignment)在分析 DNA 序列的相關性上是一重要的技術，藉由 DNA 序列的範本或圖形之共同性，可以辨認功能上的相關性，並且可以進行構造的預測及將分子層次的功能性予以模式化。

多序列比對分析程式，使用動態的變異分析方法，可以正確地測量比對的品質，而比對品質是由殘基配對間的排列方式或殘基排列間的缺口(gap)存在與否所構成，合併進行運算，以求得比對的最小值。在眞核生物基因的預測上，基因搜尋程式的預測結果證明其具有一定的精確性。

四、序列相似性搜尋(Sequence Similarity Searching)

在計算分子生物學中序列相似性搜尋是一極重要的方法，因爲從分子間相同性的比對結果，可提昇至分子序列結構性或功能性的研究。基因組資料庫的搜尋結果顯示生物上重要序列的相關性，並可擬出未來研究的策略。分子序列資料庫的相同性比對能力受到五種因素影響，即運算(algorithms)、評分系統(scoring system)、比對統計學(algorithms statistics)、資料庫更新(database updates)、資料庫序列偏歪(database sequence bias)。資料庫搜尋程式是根據局部序列相似度的比對結果而來。運算方法必須平衡速度、硬體需求及生物親緣關係等競爭因子。依照特定評分系統的計算結果進行比對，而適當評分系統的使用原則是值得深思的問題。對比對的統計結果應抱有明確的懷疑，因爲資料庫搜尋程式的運算方法並非完美。因此，如何訂定統計上顯著的估值，是非常困難的。要利用正確而廣泛的序列資料庫必須進行全面性的搜尋，但選擇時所產生的偏歪影響，仍會使誤差存在於分子序列的資料庫中。所以使用資料庫搜尋程式進行目標序列與資料庫所有序列之間的配對比較時，必須要在配對相似度上設下門檻。常用的資料庫搜尋程式主要有Smith-Waterman、FASTA及BLAST等三種。

Smith-Waterman程式[http://decypher2.stanford.edu/algo-sw/SW_nn.html-ssi]使用的是動態程式，以高敏感的相似度進行比對計算。然而，這種最適用的計算工具需要執行於二次方程式(quadratic time)。FASTA [www-nbrf.georgetown.edu/pirwww/search/fasta.html]是一種近似啓發性程式，爲一級方程式的配對比較。動態程式可用於稱爲熱點(hot spots)的一連串比對計算，此一計算爲最大序列近似度比對及全球資料庫相似度評分的組合。雖然FASTA的精準度不如Smith-Waterman程式，但FASTA的執行速度較快，因此使用者必須在精準度及執行速度中做取捨。BLAST (basic local alignment search tool) [www.ncbi.nlm.nih.gov/BLAST]是另一種近似啓發性演算式，以進行一級配對比較。BLAST所使用的熱點策略，是使

用更嚴格的限制來找出較少但較佳的比對熱點，此種策略著重於找尋不具缺口之局部高相似度區域的線性比對。由於熱點會延伸至周圍區域，即使經由數個局部相似區域鍊合所產生的缺口，對運算沒有影響。BLAST與FASTA比較具有下列優點：執行速度快、包含電腦輸出解答以及比對結果都附有統計的估值。因此，BLAST已成為目前序列資料庫的最受歡迎的搜尋工具。最近我國工研院生醫中心亦發展一套與BLAST相似，但運算速度更快的FLAG (fast local alignment for gigabases) [http://flag.itri.org.tw/]運算系統。

五、基因表現(Gene Expression)

基因表現研究的涵義為當基因表現時進行mRNA的定量測量，以顯示生物作用的特徵及闡明基因轉錄的機制。特別在受藥物及不穩定疾病影響的情況下，基因表現的研究為檢測mRNA，鑑定屬於何種基因表現是分子遺傳重要的研究課題。基因表現的高效能分析技術共有：表現序列標記(expressed sequence tags, ESTs)、DNA微陣列(DNA microarrays)、扣除選殖(subtractive cloning)、差異表現(differential display)及基因表現連續分析法(serial analysis of gene expression, SAGE)等五種。高效能基因表現分析實驗有四種目的，即新基因的鑑定、病理學作用分子標誌的鑑定、潛在藥物標籤的鑑定以及就藥物遺傳學方面說明藥物治療的分子層次。高效能基因表現分析可以同步地觀察數以千計的基因以及大量的基因表現資料。進行大規模的基因表現研究有助於構造性遺傳圖譜在功能活性區域資料的補充。因此，由靜態構造性基因組資料至動態功能性染色體組資料的研究，對基因表現的資料是一個不可或缺的項目。

基因表現資料庫包含13種資料範圍，如基因表現分析、資料庫範圍、基因表現資料(包括：基因名稱、分析方法、時間的訊息、空間的訊息、定量、基因產物、資料使用者的解析、連結記載、與其他資料的連結)、網路使用許可、網路使用規定。基因表現資料庫提供整合性的資料管理及分析系統，可用於大規模基因表現實驗所得的轉錄表現資料上。但基因表現資料庫不必為基因表現實驗資料的收集、貯存、修正以及為疑問進行明確的定義。目前人類基因表現資料庫已經被再次檢閱。

六、基因表現資料庫的開發(Gene Expression Database Mining)

目前基因表現資料庫的開發是在確認基因表現資料的本質及相關性，有助於透視基因表現的狀態以及鑑定多基因(multiple gene)的表現，例如病理學狀態的研究。基因表現資料分析使用兩種研究方法：假設檢定及知識發現。假設檢定乃是研究生物作用中的誘導或無法預期的結果。知識發現即在發現生物資料的內部結構。知識發現提供兩種方法以進行基因表現資料分析：統計(例如群聚分析(cluster analysis)及摹擬化。資料摹擬可快速描述大型基因表現資料之群聚分析結果。

七、蛋白質組學(Proteomics)

蛋白質組學的研究目的為進行藥物或疾病的影響下之蛋白質表現的定量分析。因此，蛋白質組學的研究採用蛋白質定量分析，以測定基因的表現，並利於描述生物作

用的特徵以及說明基因轉譯作用的機制。混合蛋白質的蛋白質組學分析有三個程序，蛋白質分析、蛋白質鑑定及蛋白質定量。蛋白質分析是使用二向式聚丙烯醯胺凝膠電泳(two-dimensional polyacrylamide gel electrophoresis)進行分析。蛋白質鑑定是以Edman解降法(edman degradation)、質譜儀(mass spectronmetry)及西方墨漬法(western immunoblotting)進行分析。蛋白質定量可以藉由掃瞄器及磷酸影像器(phosphorimager)的幫助完成。觀察基因的表現可以監看基因轉錄作用，而觀察蛋白質組的表現可以監看基因轉譯作用。由於受限於額外的二級轉譯作用，蛋白質組比基因表現有更多的限制及後轉譯的修飾作用(post-translational modification)。在目標基因表現型的研究上，蛋白質組顯得比基因表現更具說服力。因此，蛋白質組學對於功能性基因組資料庫提供比基因表現更直接的證據。

八、蛋白質組資料庫(Proteome Databases)

蛋白質組資料庫建構在數以千計基因之蛋白質產物的表現程度及特性，它提供整合性的資料管理及轉譯表現資料的分析系統，並且提供基因表現全面性研究的方法。蛋白質組資料庫的五大研究課題，如蛋白質產物豐富的關聯性、後轉譯的修飾作用、次細胞的定位、分子層次的轉移及蛋白質間的交互作用，並不能藉由DNA分析解決。因此，需要創造基因及基因產物的全面性資料庫，以便建構更深入作用機制的全面性資料庫，例如基因表現的調節作用、代謝作用的途徑及訊息的串連(signalling cascades)。

九、蛋白質組資料庫的開發(Proteome Database Mining)

蛋白質組資料庫開發的目的在於確認蛋白質組資料的本質及相關性。藉由分析複雜的蛋白質組資料庫可獲得兩個好處，即透視基因轉譯作用及後轉譯修飾作用以及鑑定複雜生物作用中複合蛋白質的表現，例如病理學狀態。目前蛋白質組資料庫的開發已經在三種實驗系統中使用：*E. coli* K-12蛋白組、人類淋巴蛋白質及藥物的毒性評估。

人類基因組的開發創造了兼具構造性及功能性的高解析度遺傳及物理圖譜，並產生了龐大數據資料，且從有限的資料中獲得有用的資訊。爲了便於管理及分析遺傳資料，而劃分出三種有效的資料分析層次，分別爲基因組學(genomics)、基因表現(gene expression)及蛋白質組學(proteomics)。基因組資料庫的開發可以驗證基因組轉譯爲蛋白質的區域，並利用已知功能的基因序列進行相似性比較，以便尋找其它未知基因的功能；而開發基因表現資料庫可以驗證基因轉錄作用的模式及相關性；開發蛋白質組資料庫可以驗證基因轉譯作用的模式及相關性。當人類基因組的DNA定序工作完成時，人類基因組開發的重點就會由靜態的構造性基因組學轉移至動態的功能性基因組學，今後基因表現及蛋白質組學的研究將成爲資料庫開發的主力。若基因組、基因表現及蛋白質組資料庫的開發彼此之間互補長短，將使資料庫的分析能力有更大的含蓋範圍，進而使得資料庫之序列比對分析得以改善，並可準確地預測蛋白質的功能及作用，例如基因表現、代謝途徑及訊息串連，而達成整合基因型及表現型製作完整圖譜的最終目的。